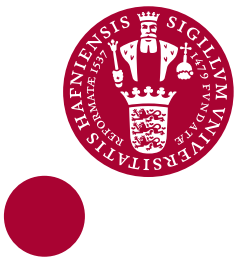

Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations

Bottaro, Sandro; Bussi, Giovanni; Kennedy, Scott D.; Turner, Douglas H.; Lindorff-Larsen, Kresten

Published in:
Science Advances

DOI:
10.1126/sciadv.aar8521

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

BIOCHEMISTRY

Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations

Sandro Bottaro,[1]* Giovanni Bussi,[2] Scott D. Kennedy,[3] Douglas H. Turner,[4] Kresten Lindorff-Larsen[1]*

RNA molecules are key players in numerous cellular processes and are characterized by a complex relationship between structure, dynamics, and function. Despite their apparent simplicity, RNA oligonucleotides are very flexible molecules, and understanding their internal dynamics is particularly challenging using experimental data alone. We show how to reconstruct the conformational ensemble of four RNA tetranucleotides by combining atomistic molecular dynamics simulations with nuclear magnetic resonance spectroscopy data. The goal is achieved by reweighting simulations using a maximum entropy/Bayesian approach. In this way, we overcome problems of current simulation methods, as well as in interpreting ensemble- and time-averaged experimental data. We determine the populations of different conformational states by considering several nuclear magnetic resonance parameters and point toward properties that are not captured by state-of-the-art molecular force fields. Although our approach is applied on a set of model systems, it is fully general and may be used to study the conformational dynamics of flexible biomolecules and to detect inaccuracies in molecular dynamics force fields.

INTRODUCTION

Many biomolecules are highly dynamic systems that undergo significant conformational rearrangements during their function. Experimental techniques such as nuclear magnetic resonance (NMR) spectroscopy, fluorescence spectroscopy, and small-angle x-ray scattering are well suited to probe the dynamics of molecules in solution. However, obtaining a full description of structure and dynamics of biomolecules using experiments alone can be highly nontrivial because the measured quantities are generally time and ensemble averages over conformationally heterogeneous states.

In this perspective, maximum entropy (*1*–*3*) (MaxEnt) and Bayesian (*4*–*6*) approaches have emerged as powerful theoretical tools for integrating simulations with experiments. These approaches typically generate a structural ensemble for the system of interest using molecular dynamics (MD) or Monte Carlo simulations. This ensemble, however, may not necessarily agree with available experimental data because of limited sampling or inaccuracies in the used model describing the physics and chemistry of the system (that is, the force field). The underlying idea behind MaxEnt is to minimally perturb a simulation ensemble to match the experimental data. Random and systematic errors can be taken explicitly into account. These approaches have been successfully used to study protein systems (*6*), whereas applications to nucleic acids have been so far limited (*7*, *8*).

Here, we consider the conformational ensembles of four RNA tetranucleotides by integrating available NMR data (*9*–*11*) with extensive atomistic MD simulations. Despite the lack of a biological relevance, RNA tetranucleotides serve as challenging model systems both from the experimental and computational point of view. First, they display significant dynamics: Therefore, one single structure cannot be representative of the entire ensemble. The conformational heterogeneity makes it nontrivial to provide a structural interpretation of average measurements using standard three-dimensional structure determination tools. Second, current state-of-the-art MD force fields fail in predicting the properties of these tetranucleotides (*12*). Several studies (*11*, *13*) have shown MD simulations to overstabilize so-called intercalated conformations (see Fig. 1) that, in some cases, correspond to the predicted free-energy minimum. From the experimental point of view, the presence and the population of intercalated conformations are expected to be low but cannot be accurately quantified.

Here, we show that, even with the aforementioned complications, it is possible to obtain an accurate thermodynamic description for a system of interest by combining experiments and simulations. We report extensive atomistic MD simulations in explicit water for r(AAAA), r(CCCC), r(GACC), and r(UUUU) tetranucleotides. We show substantial disagreement between predicted and experimental NMR data, even when using recent force-field parameters. We therefore use the MaxEnt/Bayesian approach to refine the simulated ensembles to match a set of available NMR experimental data, including nuclear Overhauser effect (NOE) intensities and scalar couplings.

Analysis of the optimal ensembles shows that r(CCCC) and r(GACC) are ≈ 60% in A-form–like conformations. r(AAAA) and r(UUUU) display a higher complexity because the optimal ensembles consist of a mixture of A-form with other conformationally heterogeneous structures.

RESULTS

Agreement between experiments and simulations

We first consider the tetranucleotide with sequence CCCC. Previous NOE measurements for r(CCCC) were found to be consistent with a conformational ensemble mostly composed of A-form–like structures, with a minor population (13%) of conformations with cytosine at position 4 (C4) inverted (see Fig. 1A) (*10*). Extensive MD simulations with the standard Assisted Model Building with Energy Refinement (AMBER) force field (χ_{OL3} described in Materials and Methods) showed the presence of highly populated intercalated structures in which C1 is interposed between C3 and C4 (*11*, *13*), whereas C2 is either stacked on C3 or solvent-exposed. The lack of A-form–like structures is confirmed in our χ_{OL3} simulations, as shown in the eRMSD histogram from an ideal A-form in Fig. 1B (yellow line). To measure distances between three-dimensional structures, we here use the eRMSD, an RNA-specific

[1]Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [2]International School for Advanced Studies, Trieste, Italy. [3]Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, NY 14642, USA. [4]Department of Chemistry, University of Rochester, Rochester, NY 14627, USA.
*Corresponding author. Email: sandro.bottaro@bio.ku.dk (S.B.); lindorff@bio.ku.dk (K.L.-L.)

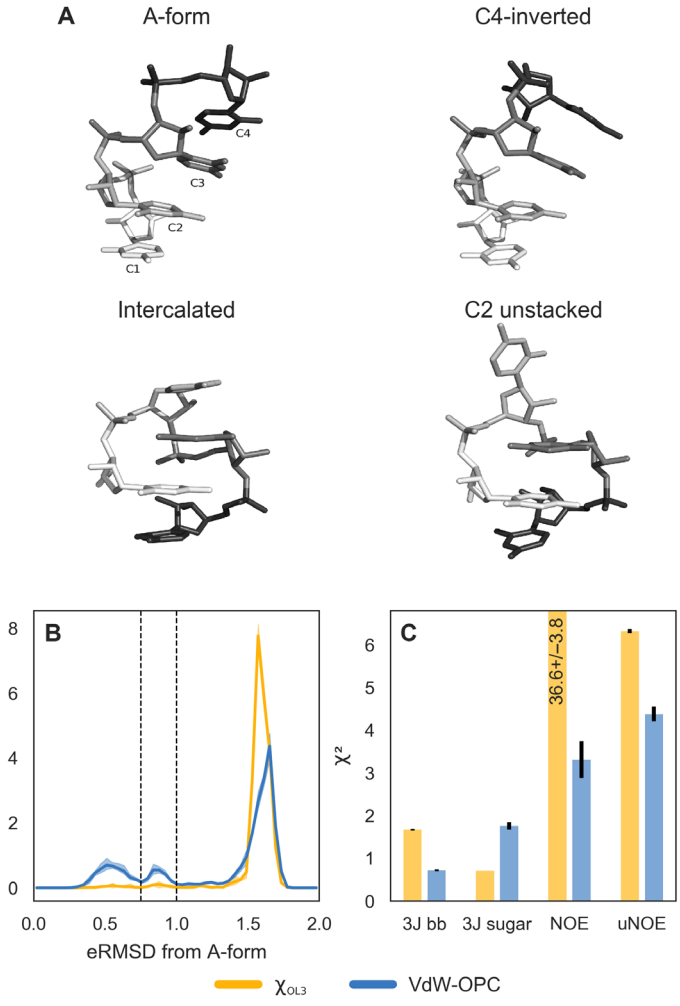

Fig. 1. Conformational ensemble of r(CCCC) simulations and agreement with experimental data. (**A**) Three-dimensional structures of r(CCCC) discussed in the main text. (**B**) eRMSD from the A-form histogram for χ_{OL3} and VdW-OPC simulations. Solid lines indicate the average calculated using a blocking procedure, whereas the area between minimum and maximum is shown in shade. The histogram displays three peaks corresponding to different conformations: A-form–like (eRMSD < 0.75), C4-inverted (0.75 < eRMSD < 1), and intercalated/C2 unstacked (eRMSD > 1.0). Thresholds are shown as dashed lines. (**C**) Agreement between simulations and experiments quantified using the χ^2 statistic for backbone scalar couplings (3J bb), sugar scalar couplings, NOE, and uNOE. Error bars in black show the SEM. The value of χ^2 relative to NOE with χ_{OL3} is out of scale; the corresponding value with error is therefore reported in the figure.

metric distance based on the relative orientation and position of nucleobases (*14*). It has recently been reported (*15*) that corrections to oxygen van der Waals radii (*16*) in conjunction with the optimal 3-charge, 4 point (OPC) water model (*17*) (hereafter referred to as VdW-OPC) significantly disfavor the presence of intercalated structures in r(GACC) and r(CCCC) tetranucleotides, thereby stabilizing A-form–like conformations. When using the VdW-OPC force field (Fig. 1B, blue line), we observe a small, yet significant population of A-form–like structures (eRMSD < 0.75) and C4-inverted conformations (0.75 to 1.0 eRMSD from A-form).

The higher accuracy of VdW-OPC with respect to χ_{OL3} is further confirmed by the improved agreement between calculated and experimental data. Figure 1C reports the χ^2 for backbone ^{3}J scalar couplings (H3-P, H5′/H5″-P, and H4-H5′/H5″), sugar ^{3}J couplings (H1′-H2′, H2′-H3′, and H3′-H4′), and NOE intensities (*10*, *11*). In addition, we consider the absence of specific peaks in NOE spectroscopy (NOESY) data as a source of information. On the basis of assigned chemical shifts, NMR spectra were inspected for the presence of NOE cross peaks between every pair of nonexchangeable protons in the tetramers. To assign unobserved NOEs (uNOEs), we estimated the maximum NMR observable distance for each potential NOE from the minimum detectable cross-peak volume (see Materials and Methods). Whenever simulations predict a shorter distance between these proton pairs, it is considered a violation of a uNOE. Note that the importance of uNOE has been discussed for protein systems as well (*18*). uNOEs are of particular importance because several violations are present in intercalated structures (*11*). It can be clearly seen in Fig. 1C that the VdW-OPC force field provides a better agreement with experimental data, especially for NOEs. We note, however, the higher χ^2 for ^{3}J sugar scalar couplings with respect to the standard χ_{OL3} force field.

Reweighting procedure

It is evident from Fig. 1C that the conformational ensemble predicted by simulations alone is not in complete agreement with experiments. We therefore generate a conformational ensemble that satisfies the experimental constraints using the MaxEnt/Bayesian approach with the inclusion of error treatment (*5*, *7*). In MaxEnt approaches, one seeks the minimal perturbation of the simulated ensemble (that is, the prior distribution) that satisfies a set of known experimental averages. This can be achieved (*2*, *7*) by minimizing the function

$$\Gamma = \log(Z(\lambda)) + \sum_i^m \lambda_i F_i^{\mathrm{EXP}} + \frac{1}{2}\sum_i^m \lambda_i^2 \sigma_i^2 \quad (1)$$

with respect to the set of Lagrange multipliers $\lambda = \lambda_1 \ldots \lambda_m$. Here, the index i runs over the m experimental averages F_i^{EXP} with associated normally distributed and uncorrelated errors σ_i. Z is the partition function $Z(\lambda) = \sum_j^N w_j^0 \exp[-\sum_i^m \lambda_i F_i(\mathbf{x}_j)]$, where $F_i(\mathbf{x}_j)$ is the function used to back-calculate the experimental observable from the atomic coordinates $\mathbf{x}$, and $\{w_1^0 \ldots w_N^0\}$ corresponds to the weights of the N frames in the prior distribution. Note that this approach is completely equivalent to a Bayesian ensemble refinement approach (*5*, *19*) in which one seeks the optimal weights $\{w_1 \ldots w_N\}$ minimizing the negative log posterior L

$$L(w_1 \ldots w_N) = \frac{m}{2}\chi^2 + \theta S_{\mathrm{REL}} \quad (2)$$

where $\chi^2 = \sum_i^m (\sum_j^N w_j F_i(x_j) - F_i^{\mathrm{EXP}})^2 / m\sigma_i^2$ is the deviation from the experimental averages, and the relative entropy $S_{\mathrm{REL}} = \sum_j^N w_j \log(w_j / w_j^0)$ quantifies the deviation from the prior distribution. θ sets the relative weight between these two quantities and needs to be chosen by considering how χ^2 and S_{REL} vary for different values of this parameter (*5*), as described below.

A few items are worth highlighting. First, the number of experimental constraints, m, is typically much smaller compared to the number of samples, N, and it is therefore in practice easier to minimize the function in Eq. 1 rather than Eq. 2. Second, θ enters the MaxEnt formulation (Eq. 1) as a global scaling factor of all Gaussian errors σ_i. Third, heterogeneous data (NOE, ^{3}J couplings, chemical shifts, etc.)

can be used simultaneously in the reweighting procedure, both averages and inequality constraints (*7*).

Choosing the data and the confidence parameter

Before proceeding to the analysis of the optimized ensemble, we study the dependence of the results on (i) the type of experimental data used for reweighting and (ii) the tunable parameter θ. Given the better initial agreement with experimental data, we here consider the VdW-OPC simulations. Figure 2A (solid lines) shows χ^2 as a function of θ when using scalar couplings as the only input for reweighting. As expected, small θ corresponds to a better fit, whereas in the limit of large θ we approach the original, unreweighted χ^2 value (dotted-dashed line). We can also monitor the behavior of χ^2 relative to data that were not used in the reweighting (Fig. 2A, dashed line). In the limit of $\theta \to 0$, the violations of uNOE become very small. Conversely, the agreement with NOE distances has a clear minimum around θ = 3. When using only NOEs for reweighting (Fig. 2B), we observe improved agreement with respect to all other experimental sources of data. This effect is more pronounced when using uNOE only (Fig. 2C), demonstrating the importance and the validity of this type of data. Note that, at least for r(CCCC), the reweighted χ^2 values are always smaller compared to the original, unreweighted values, indicating that the different types of data are consistent. Given the cooperative effect of the different types of data, we finally consider the case in which ^{3}J couplings, NOE, and uNOE are all used at the same time for reweighting (Fig. 2D). This combination provides the best accord both for r(CCCC) and for the other tetranucleotides (figs. S1 to S3).

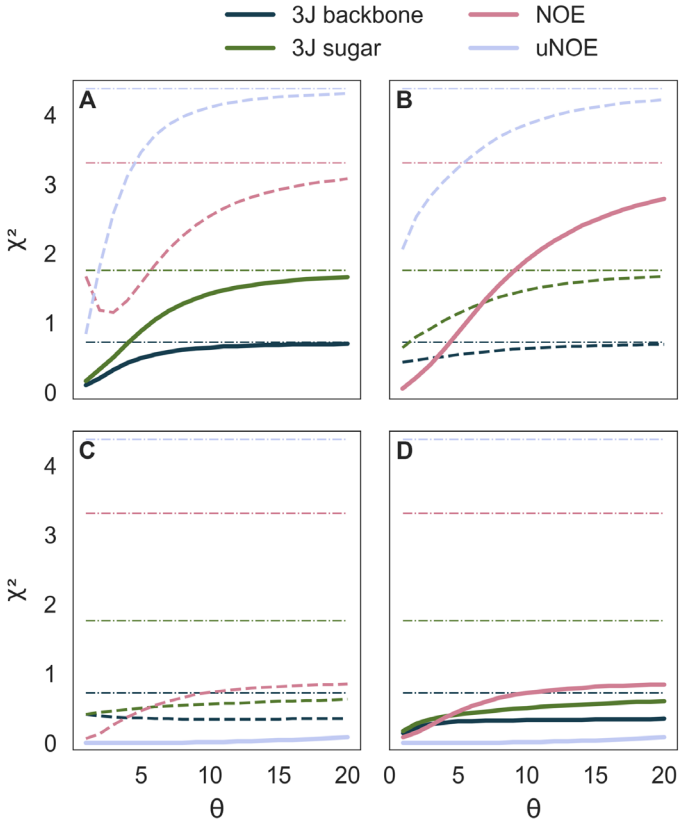

Fig. 2. Agreement between reweighted r(CCCC) simulations and experiments using different data for reweighting (solid lines) and for validation (dashed lines). (**A**) χ^2 as a function of θ when using the two sets of ^{3}J scalar couplings for reweighting and cross-validated against NOE and uNOE. Results using NOE distances and uNOE distances for reweighting are shown in (**B**) and (**C**), respectively. (**D**) Results using all three types of data for reweighting. Initial, unreweighted χ^2 are shown as dotted-dashed lines. Values of χ^2 below 1 correspond to an average difference between prediction and experiment within the error σ.

When considering χ^2 alone, one would choose a small θ so as to attain the best fit. In the limit $\theta \to 0$, however, the original ensemble can be substantially distorted to the point that the physicochemical information contained in the force field is lost (Eq. 2). In addition, this has a detrimental effect on the statistical errors because the number of effective frames contributing to the ensemble decreases significantly (fig. S4). To strike a good balance between fit and proximity to the prior distribution, we scan different values of θ until a further decrease of this parameter leads to an increase in the relative entropy without substantially improving the fit (*5*). Although this procedure does not provide a unique θ, it makes it possible to identify a range of reasonable values (fig. S4). We here use a pragmatic approach and set θ = 2, the largest value for which $\chi^2 < 2$ for all tetranucleotides and all types of experimental data. Note that the relative weight of different experiments might be modulated by changing the corresponding values of σ. Scatter plots comparing individual experimental averages against simulations before/after reweighting are shown in figs. S5 to S8.

Conformational ensemble of r(CCCC)

The set of optimized weights can be now used to calculate the full probability distribution of any observable (for example, distances, torsion angles, etc.). To appreciate the properties of the optimized ensemble, it is again interesting to consider the distribution of the distance from A-form (Fig. 3A).

The original VdW-OPC MD ensemble consists of ≈18% A-form structures (eRMSD from A-form < 0.75) and 9% with C4 either inverted or unstacked (eRMSD from A-form in the 0.75 to 1.0 range). From the histogram of eRMSD relative to intercalated structure (Fig. 3B), the initial ensemble estimates a 53% population of intercalated structures that can be subdivided into fully stacked intercalation (13%, eRMSD < 0.4) and intercalated structures with C2 unstacked (≈ 40%, eRMSD in the 0.4 to 0.8 range).

Upon reweighting, A-form represents the major conformation (54%) followed by C4 inverted (22%). The population of intercalated structures is significantly reduced in the reweighted ensemble to ≈7% (Fig. 3B). This result is not surprising because it is consistent with the picture proposed in the original experimental paper (*10*). The ensemble obtained here, however, did not require expert interpretation of the individual NOE distances. The reweighting approach takes into account general properties encoded in the force field and makes it possible to monitor degrees of freedom that were not measured by NMR. Two significant examples are reported in Fig. 3 (C and D). Figure 3C shows the distribution of the distance between the atom OP2 in C3 and the hydrogen at the 5′ terminus in C1 (H5T), where we observe the presence of a stable hydrogen bond between these two atoms (associated with the intercalated conformation) that is almost absent after reweighting. The reweighting also markedly affects the distribution of the α angle in C2, because we find that gauche$^-$ (g^-) is the preferred rotameric state in the reweighted ensemble (Fig. 3D). A similar behavior is observed for α in C3 and ζ in C2 and in C3, in accordance with previous simulation studies that have shown the importance of these two torsion angles in tetranucleotides and tetraloop simulations (*20*, *21*). We highlight that the backbone

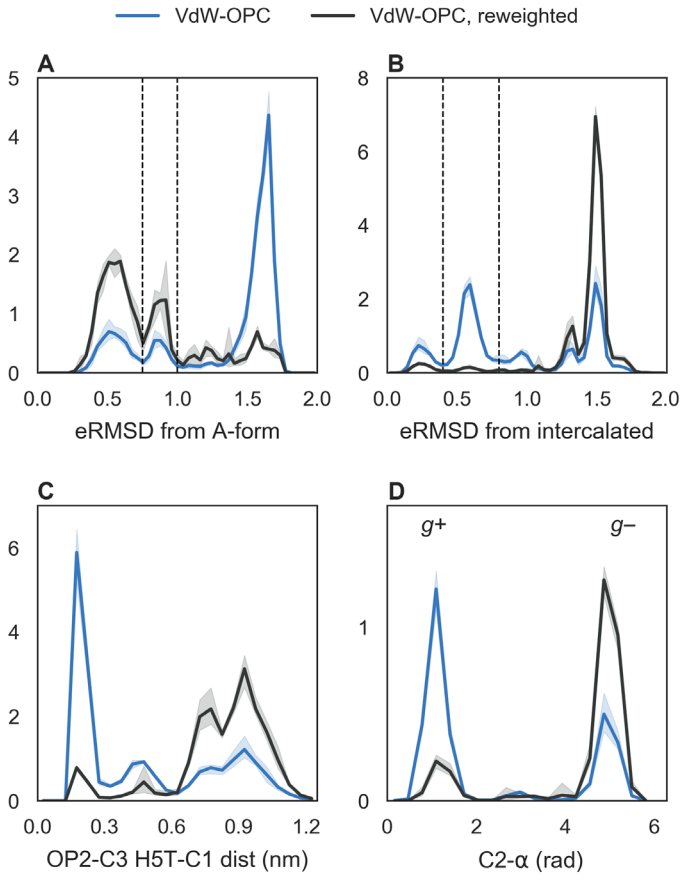

Fig. 3. Distribution of different observables before and after reweighting r(CCCC) simulations using VdW-OPC. Solid lines indicate the average calculated using a blocking procedure; minima and maxima are shown in shade. (**A**) eRMSD from ideal A-form, (**B**) eRMSD from an intercalated conformation, (**C**) distance distribution between OP2 in C3 and H5T in C1, and (**D**) the α torsion angle of C2. Peaks in (A) and (B) can be associated to the structures shown in Fig. 1: A-form (eRMSD from A-form below 0.75), C4-inverted (eRMSD from A-form 0.75 to 1.0), intercalated (eRMSD from intercalated <0.4), and intercalated with C2 unstacked (eRMSD from intercalated 0.4 to 0.8). eRMSD boundaries are shown as dashed lines.

^{3}J scalar couplings used in the reweighting procedure report on ϵ and γ angles, but not on α/ζ.

Conformational ensemble of r(AAAA), r(GACC), and r(UUUU)

The same procedure described above was applied to r(AAAA), r(GACC), and r(UUUU) tetranucleotides. In all cases, VdW-OPC is considerably better compared with the χ_{OL3} force field (Fig. 4, left panels). The reweighting procedure further improves agreement with experimental data. However, we do observe a residual discrepancy in some cases ($\chi^2 > 1$) that stems from predicted NOE distances falling outside the experimental range (figs. S5 to S8). In the case of r(GACC), three NOEs reported in the original experimental work (*11*) were not satisfied in a preliminary reweighting. After careful checking of the experimental data, we discovered two previously undetected spectral overlaps. The corresponding NOEs were thus removed from the list of data points. Evidently, the reweighting procedure can be used to highlight data points that are inconsistent with the others and hence might require manual inspection. These cases can be treated by using error models suitable to describe outliers (*7*, *22*).

The r(AAAA) ensemble is composed of ≈30% A-form–like structures and 16% A4-inverted/unstacked (Fig. 4, middle panels). In this case, the available experimental data could not completely rule out the presence of intercalated structures, which represent the 13% of the optimized ensemble (Fig. 4, right panels). The remaining 40% is composed of other structures that exhibit one or more sugar puckers in C2′-endo and/or the A1-χ angle in syn conformation (Table 1 and fig. S9).

r(GACC) behaves very similarly to r(CCCC), with ≈ 60% A-form–like structures and 20% C4-inverted/unstacked. The similarity between r(GACC) and r(CCCC) can also be appreciated by considering the sugar pucker and χ angle preferences reported in Table 1 and figs. S10 and S11. Intercalation is almost completely absent in the reweighted ensembles.

Among all the systems studied here, r(UUUU) has the lowest population of A-form–like structures (9%). The rest of the ensemble is composed of a variety of diverse structures that cannot be easily clustered. This can be seen from the low percentage of sugar pucker in C3′-endo conformation (Table 1 and fig. S12) and from the relatively flat distribution of eRMSD from A-form in Fig. 4. Among this set of diverse conformations, a very small fraction of intercalated structures are present.

Note that the percentages reported here depend on two important choices: on the reference structures and on the choice of θ. Whereas the geometry of the ideal A-form can be unambiguously defined (*23*), the intercalated structures are obtained by performing a cluster analysis of the χ_{OL3} simulation as described previously (*24*). Although this choice has a degree of arbitrariness, we found it as a useful and intuitive manner to define an order parameter complementary to the distance from A-form. As for θ, we verified that the population of the different states do not depend critically on this parameter in the relevant range $2 < \theta < 5$ (fig. S13).

DISCUSSION

Here, we have described the structural ensembles of four RNA tetranucleotides at the atomistic level. The characterization of these systems represents a first step in understanding the ensembles and internal dynamics of larger oligonucleotides and other RNA molecules undergoing significant conformational changes. Despite their apparent simplicity, tetranucleotides are particularly challenging systems: Because of their conformational heterogeneity, NMR experimental data need to be interpreted as ensemble averages. For this reason, standard procedures for NMR structure determination cannot be easily applied (*25*). In addition, it is not possible to predict the properties of these systems using simulations alone, because of known force-field inaccuracies (Fig. 1). Only the combination of experiment with computation makes it possible to provide an atomic-detailed description of their conformational ensembles. In this context, the MaxEnt/Bayesian approach serves as a fundamental theoretical ingredient for using the two techniques in conjunction.

We find that r(CCCC) and r(GACC) are ≈ 60% in A-form–like conformations and ≈ 20% with the 3′ terminal base either unstacked or inverted (Fig. 1A). r(AAAA) tetranucleotide is characterized by a lower A-form content (≈ 30%) and displays a larger variability in terms of sugar conformations. Our analysis shows that the presence of intercalated structures cannot be excluded in this case. Among the four systems considered here, r(UUUU) displays the highest disorder (Table 2), with a percentage of A-form conformation of ≈10%.

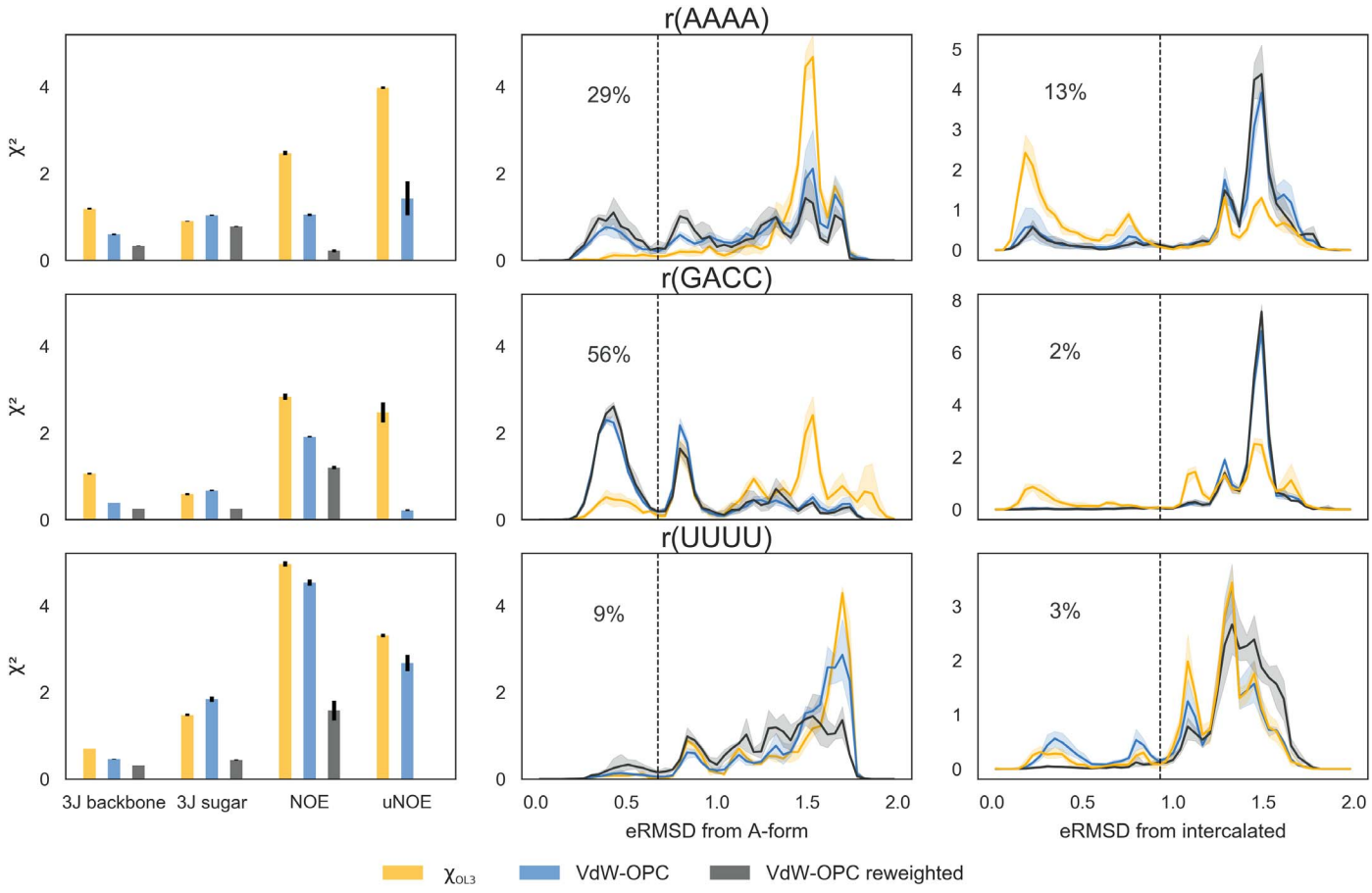

Fig. 4. Comparison between reweighted and unreweighted ensembles for r(AAAA), r(GACC), and r(UUUU) tetranucleotides. (**Left**) Agreements between calculated and experimental averages for χ_{OL3}, VdW-OPC, and reweighted VdW-OPC simulations. (**Middle**) Histograms of eRMSD from ideal A-form. (**Right**) Histograms of eRMSD from intercalated structure. The dashed lines indicate the thresholds used for calculating the percentage of A-form–like (middle) or intercalated structures (right) upon reweighing.

Table 1. Percentage of C3′-endo (δ < 115°) and anti (χ > 120°) of reweighted VdW-OPC simulations. The statistical error calculated using block averaging is below 1%.

	Sequence	N1	N2	N3	N4
% C3′-endo	AAAA	70.6	75.1	84.5	66.3
	CCCC	90.7	88.9	88.5	71.7
	GACC	86.9	87.1	88.3	71.1
	UUUU	61.4	49.5	50.5	63.2
% Anti	AAAA	65.2	96.5	98.4	97.8
	CCCC	98.0	98.5	99.8	99.7
	GACC	89.2	99.9	98.9	99.4
	UUUU	88.5	97.1	96.8	96.3

Table 2. Number of experimental averages.

	NOE	^{3}J sugar	^{3}J backbone	uNOE
AAAA	36	11	17	243
CCCC	27	11	15	245
GACC	20	12	17	284
UUUU	9	10	15	282

From a technical perspective, the combination of experiments and simulations can be seen as a regularization problem in which a small set of experimental data is used to gain insights into a highly dimensional, complex set of molecular conformations. The problem is underdetermined and has to be regularized by using a suitable prior distribution, here provided by MD simulations. This interpretation becomes transparent in the Bayesian ensemble refinement formulation in Eq. 2 (*5*, *19*). The balance between fit quality (χ^2) and deviation from the prior distribution (S_{REL}) is tuned by a system-dependent, global confidence parameter θ, that is not known a priori.

In a number of recent MaxEnt-inspired approaches, a bias deriving from the experimental data is estimated on the fly during the simulations (*5*, *7*, *22*, *26*). These approaches have the advantage of enhancing the sampling in relevant regions of the conformational space. On the other hand, the reweighting procedure can be applied a posteriori to

existing simulations whenever new experimental data are available (*27*). Because reweighting only requires a cheap post-processing of existing trajectories, it is straightforward to perform multiple cross-validation tests. In addition, reweighting is very convenient when the forward model calculation is particularly demanding, because in biased methods the back calculation of averages from structures has to be performed at least every few time steps (*28*).

Here, we have found that combining experimental data with simulations had mutual beneficial effects. On the one hand, simulations helped identifying spurious experimental data points. On the other hand, we have used experimental data to identify inaccuracies in MD force fields. Modern atomistic force fields consist of hundreds of parameters, and even finding the relevant interactions that can potentially improve their accuracy is a time-consuming and nontrivial task. Our approach substantially simplifies this search (Fig. 3, C and D), because the probability distribution over any degree of freedom before and after reweighting can be readily compared. We find that hydrogen bonds to nonbridging oxygens are significantly destabilized upon reweighting, in accordance with previous simulation studies (*11*, *29*). At the same time, the population of α and γ torsion angles is, in some cases, shifted from gauche$^+$ to gauche$^-$. As molecular mechanics force fields improve, the approach described here should require less experimental data to provide reliable determination of structural ensembles (*30*, *31*).

MATERIALS AND METHODS

MD simulations

We performed MD simulations on r(AAAA), r(CCCC), r(UUUU), and r(GACC) tetranucleotides. Each system was simulated with two different force fields: (i) the AMBER 99 force field (*32*) with parmbsc0 corrections to α/γ (*33*) and the χOL corrections to χ torsion angles (*34*) in TIP3P water. We refer to this combination as χ_{OL3}. These simulations were taken from our previous studies (*20*, *35*). (ii) χ_{OL3} with corrections to van der Waals oxygen radii (*16*) (atom types O2, OH, and OS) and using the OPC water model (*17*). We refer to this combination as VdW-OPC. Parameters are available at http://github.com/srnas/ff. MD simulations were performed using the GROningen MAchine for Chemical Simulations (GROMACS) 4.6.7 software package (*36*). Ideal A-form, fully stacked initial conformations were generated using the Make-NA web server. The oligonucleotides were solvated in a truncated dodecahedric box and neutralized by adding Na^+ counterions (*37*). Initial conformations were minimized in vacuum first, followed by a minimization in water and equilibration in NPT ensemble at 300 K and 1 bar for 1 ns. Production runs were performed in the canonical ensemble using a stochastic velocity rescaling thermostat (*38*). All bonds were constrained with the Linear Constraint Solver algorithm, and equations of motion were integrated with a time step of 2 fs. Tetranucleotides were simulated using temperature replica exchange (*39*) using 24 replicas in the temperature range of 278 to 400 K for 1.0 μs per replica. All the analyses presented here were performed for the 300 K replica and using 20,000 frames. Averages and SEMs were calculated using four blocks of 5000 samples each. Sampling was sufficient to achieve similar eRMSD distributions for each block (fig. S14) and to obtain populations of different substates in agreement with multidimensional replica exchange MD simulations (*12*, *13*, *15*).

NMR data

Experimental NOE and scalar couplings have previously been published (*10*, *11*). We used Gaussian-distributed experimental errors of 1.5 Hz for scalar couplings and of 0.1 Å for uNOE. The error for NOE was estimated as $\min(r_{max}^{EXP} - r^{EXP}, r^{EXP} - r_{min}^{EXP})$. The number of experimental averages for each NMR parameter and for each tetranucleotide sequence is reported in Table 2. The complete list of experimental data is available in the Supplementary Materials. NOE intensities from simulations are calculated as averages over the N samples $NOE_{CALC} = (\Sigma_i^N w_i r_i^{-6})$. ^{3}J scalar couplings were calculated using the Karplus relationships described in fig. S15 and table S1 using the software baRNAba https://github.com/srnas/barnaba. Note that in some cases, the error introduced by the forward model is significant. As an example, ^{3}J scalar couplings calculated using Karplus relationships can introduce errors up to 2 Hz (fig. S16). Care should also be taken when calculating NOE intensities from proton-proton distances because the simple r^{-6} averaging does not take spin diffusion into account, and it is only valid in the limit of slow internal motion compared to the tumbling time (*40*).

Unobserved NOE

NMR spectra were inspected for the presence of NOESY cross peaks between every pair of protons in the tetramer. If no cross peak was observed, then the potential contact was classified as a uNOE. If the spectral position of a potential cross peak did not overlap any other observed cross peak, then the minimum detectable cross-peak volume was assumed to be two times the SD of spectral noise (V_{err}). Scalar coupling results in NOE cross peaks that are split into multiplets of two, four, or more peaks, resulting in accordingly reduced peak heights and increased minimum detectable volume. For a cross peak consisting of M multiplets, the minimum detectable volume is $2MV_{err}$. V_{err} and a scaling factor, c, obtained in the original work (*10*, *11*) from NOESY spectra with a 200-ms mixing time, are used to associate a distance, R, with the minimum detectable volume: $R = (c/2MV_{err})^{1/6}$. The analysis of uNOEs was carried out here with 800-ms NOESY spectra, where cross peaks are typically 2.5- to 3-fold greater than at 200 ms, so the minimum detectable NOE volume was reduced by a factor of 2.5 (after correcting for any difference in the number of NMR scans). If the spectral position of a potential cross peak partially overlapped one or more observed cross peaks, then the minimum detectable volume of the potential cross peak was determined by the magnitude of the observed cross peak and exact details of the overlap (instead of spectral noise). Typically, if the partially overlapped observed cross peak was medium or weak, respectively, then a potential cross peak exhibiting no apparent intensity was classified as unobserved with a volume that corresponded to an internuclear distance of greater than 3.3 or 4.0 A. If the overlapping observed cross peak was strong or the potential cross peak was close to the diagonal, then the potential cross peak was not classified as unobserved.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/4/5/eaar8521/DC1

fig. S1. Agreement between r(AAAA) simulations using VdW-OPC and experiments as a function of the parameter θ.

fig. S2. Agreement between r(GACC) simulations using VdW-OPC and experiments as a function of the parameter θ.

fig. S3. Agreement between r(UUUU) simulations using VdW-OPC and experiments as a function of the parameter θ.

fig. S4. χ^2 versus relative entropy and fraction of effective frames as function of θ.

fig. S5. Reweighted r(AAAA) simulations using VdW-OPC.

fig. S6. Reweighted r(CCCC) simulations using VdW-OPC.

fig. S7. Reweighted r(GACC) simulations using VdW-OPC.

REFERENCES AND NOTES

1. E. T. Jaynes Where do we stand on maximum entropy?, in *The Maximum Entropy Formalism*, R. D. Levine, M. Tribus, Eds. (MIT Press, 1978), pp. 15–118.
2. J. W. Pitera, J. D. Chodera, On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.* **8**, 3445–3451 (2012).
3. W. Boomsma, J. Ferkinghoff-Borg, K. Lindorff-Larsen, Combining experiments and simulations using the maximum entropy principle. *PLOS Comput. Biol.* **10**, e1003406 (2014).
4. D. Schneidman-Duhovny, R. Pellarin, A. Sali, Uncertainty in integrative structural modeling. *Curr. Opin. Struct. Biol.* **28**, 96–104 (2014).
5. G. Hummer, J. Köfinger, Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **143**, 243150 (2015).
6. M. Bonomi, G. T. Heller, C. Camilloni, M. Vendruscolo, Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **42**, 106–116 (2017).
7. A. Cesari, A. Gil-Ley, G. Bussi, Combining simulations and solution experiments as a paradigm for RNA force field refinement. *J. Chem. Theory Comput.* **12**, 6192–6200 (2016).
8. A. N. Borkar, M. F. Bardaro, C. Camilloni, F. A. Aprile, G. Varani, M. Vendruscolo, Structure of a low-population binding intermediate in protein-RNA recognition. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7171–7176 (2016).
9. I. Yildirim; H. A. Stern, J. D. Tubbs, S. D. Kennedy, D. H. Turner, Benchmarking AMBER force fields for RNA: Comparisons to NMR spectra for single-stranded r(GACC) are improved by revised χ torsions. *J. Phys. Chem. B* **115**, 9261–9270 (2011).
10. J. D. Tubbs, D. E. Condon, S. D. Kennedy, M. Hauser, P. C. Bevilacqua, D. H. Turner, The nuclear magnetic resonance of CCCC RNA reveals a right-handed helix, and revised parameters for AMBER force field torsions improve structural predictions from molecular dynamics. *Biochemistry* **52**, 996–1010 (2013).
11. D. E. Condon, S. D. Kennedy, B. C. Mort, R. Kierzek, I. Yildirim, D. H. Turner, Stacking in RNA: NMR of four tetramers benchmark molecular dynamics. *J. Chem. Theory Comput.* **11**, 2729–2742 (2015).
12. C. Bergonzo, N. M. Henriksen, D. R. Roe, J. M. Swails, A. E. Roitberg, T. E. Cheatham III, Multidimensional replica exchange molecular dynamics yields a converged ensemble of an RNA tetranucleotide. *J. Chem. Theory Comput.* **10**, 492–499 (2014).
13. C. Bergonzo, N. M. Henriksen, D. R. Roe, T. E. Cheatham III, Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA* **21**, 1578–1590 (2015).
14. S. Bottaro, F. Di Palma, G. Bussi, The role of nucleobase interactions in RNA structure and dynamics. *Nucleic Acids Res.* **42**, 13306–13314 (2014).
15. C. Bergonzo, T. E. Cheatham III, Improved force field parameters lead to a better description of RNA structure. *J. Chem. Theory Comput.* **11**, 3969–3972 (2015).
16. T. Steinbrecher, J. Latzer, D. A. Case, Revised AMBER parameters for bioorganic phosphates. *J. Chem. Theory Comput.* **8**, 4405–4412 (2012).
17. S. Izadi, R. Anandakrishnan, A. V. Onufriev, Building water models: A different approach. *J. Phys. Chem. Lett.* **5**, 3863–3871 (2014).
18. B. Zagrovic, W. F. van Gunsteren, Comparing atomistic simulation data with the NMR experiment: How much can NOEs actually tell us? *Proteins* **63**, 210–218 (2006).
19. B. Różycki, Y. C. Kim, G. Hummer, SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* **19**, 109–116 (2011).
20. A. Gil-Ley, S. Bottaro, G. Bussi, Empirical corrections to the Amber RNA force field with Target Metadynamics. *J. Chem. Theory Comput.* **12**, 2790–2798 (2016).
21. S. Bottaro, P. Banáš, J. Šponer, G. Bussi, Free energy landscape of GAGA and UUCG RNA tetraloops. *J. Phys. Chem. Lett.* **7**, 4032–4038 (2016).
22. M. Bonomi, C. Camilloni, A. Cavalli, M. Vendruscolo, Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2**, e1501177 (2016).
23. T. J. Macke, D. A. Case, Modeling unusual nucleic acid structures, in *Molecular Modeling of Nucleic Acids* (ACS Publications, 1998), pp 379–393.
24. S. Bottaro, K. Lindorff-Larsen, Mapping the universe of RNA tetraloop folds. *Biophys. J.* **113**, 257–267 (2017).
25. P. Sripakdeevong, M. Cevec, A. T. Chang, M. C. Erat, M. Ziegeler, Structure determination of noncanonical RNA motifs guided by ^{1}H NMR chemical shifts. *Nat. Methods* **11**, 413–416 (2014).
26. A. D. White, G. A. Voth, Efficient and minimal method to bias molecular simulations with experimental data. *J. Chem. Theory Comput.* **10**, 3023–3030 (2014).
27. S. Olsson, D. Strotz, B. Vögeli, R. Riek, A. Cavalli, The dynamic basis for signal propagation in human Pin1-WW. *Structure* **24**, 1464–1475 (2016).
28. M. J. Ferrarotti, S. Bottaro, A. Pérez-Villa, G. Bussi, Accurate multiple time step in biased molecular simulations. *J. Chem. Theory Comput.* **11**, 139–146 (2014).
29. C. Yang, M. Lim, E. Kim, Y. Pak, Predicting RNA structures via a simple van der Waals correction to an all-atom force field. *J. Chem. Theory Comput.* **13**, 395–399 (2017).
30. J. Šponer, G. Bussi, K. Miroslav, P. Banáš, S. Bottaro, R. A. Cunha, A. Gil-Ley, G. Pinamonti, S. Poblete, P. Jurečka, N. Walter, M. Otyepka, RNA structural dynamics as captured by molecular simulations: A comprehensive overview. *Chem. Rev.* 10.1021/acs.chemrev.7b00427 (2018).
31. D. Tan, S. Piana, R. M. Dirks, D. E. Shaw, RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. U.S.A.* **11**, 201713027 (2018).
32. J. Wang, P. Cieplak, P. A. Kollman, How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **21**, 1049–1074 (2000).
33. A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham III, C. A. Laughton, M. Orozco Refinement of the AMBER force field for nucleic acids: Improving the description of α γ conformers. *Biophys. J.* **92**, 3817–3829 (2007).
34. M. Zgarbová, M. Otyepka, J. Sponer, Mlaadek, P. Banas, T. E. Cheatham III, P. Jurecka, Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.* **7**, 2886–2902 (2011).
35. S. Bottaro, A. Gil-Ley, G. Bussi, RNA folding pathways in stop motion. *Nucleic Acids Res.* **44**, 5883–5891 (2016).
36. S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, E. Lindahl, GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
37. I. S. Joung, T. E. Cheatham III, Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **112**, 9020–9041 (2008).
38. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
39. Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
40. J. Tropp, Dipolar relaxation and nuclear Overhauser effects in nonrigid molecules: The effect of fluctuating internuclear distances. *J. Chem. Phys.* **72**, 6035–6043 (1980).
41. M. M. Mooren, S. S. Wijmenga, G. A. van der Marel, J. H. van Boom, C. W. Hilbers, The solution structure of the circular trinucleotide cr (GpGpGp) determined by NMR and molecular mechanics calculation. *Nucleic Acids Res.* **22**, 2658–2666 (1994).
42. C. Sich, Ohlenschläger, R. Ramachandran, M. Görlach, L. R. Brown Structure of an RNA hairpin loop with a 5′-CGUUUCG-3′ loop motif by heteronuclear NMR spectroscopy and distance geometry. *Biochem.* **36**, 13989–14002 (1997).
43. S. Nozinovic, B. Fürtig, H. R. Jonker, C. Richter, H. Schwalbe High-resolution NMR structure of an RNA model system: The 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res.* **38**, 683–694 (2010).
44. B. Reif, K. Wörner, S. Quant, J. P. Marino, J. W. Engels, C. Griesinger, H. Schwalbe, A new experiment for the measurement of nJ (C, P) coupling constants including 3J (C4i, Pi) and 3J (C4i, Pi+ 1) in oligonucleotides. *J. Biomol. NMR* **12**, 223–230 (1998).
45. C. Haasnoot, F. de Leeuw, C. Altona, The relationship between proton-proton NMR coupling constants and substituent electronegativitiesI: An empirical generalization of the Karplus equation. *Tetrahedron* **36**, 2783–2792 (1980).
46. D. B. Davies Conformations of nucleosides and nucleotides. *Prog. Nucl. Magn. Reson. Spectrosc.* **12**, 135–225 (1978).
47. J. P. Marino, H. Schwalbe, C. Griesinger, *J*-coupling restraints in RNA structure determination. *Acc. Chem. Res.* **32**, 614–623 (1999).
48. P. P. Lankhorst, C. A. Haasnoot, C. Erkelens, C. Altona, Carbon-13 NMR in conformational analysis of nucleic acid fragments 2. A reparametrization of the Karplus equation for vicinal NMR coupling constants in CCOP and HCOP fragments. *J. Biomol. Struct. Dyn.* **1**, 1387–1405 (1984).
49. C. H. Lee, R. H. Sarma, Aqueous solution conformation of rigid nucleosides and nucleotides. *J. Am. Chem. Soc.* **98**, 3541–3548 (1976).
50. J. H. Ippel, S. S. Wijmenga, R. De Jong, H. A. Heus, C. W. Hilbers, E. De Vroom, G. A. Van der Marel, J. H. Van Boom Heteronuclear scalar couplings in the bases and sugar rings of nucleic acids: Their determination and application in assignment and conformational analysis. *Magn. Reson. Chem.* **34**, S156–S176 (1996).

Acknowledgments: We thank S. Olsson for helpful comments on the manuscript. **Funding:** The research is funded by a grant from The Velux Foundations (S.B. and K.L.-L.), a Hallas-Møller Stipend from the Novo Nordisk Foundation (K.L.-L.), and the Lundbeck Foundation BRAINSTRUC initiative (K.L.-L.). G.B. has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC grant agreement no. 306662 (S-RNA-S). D.H.T. was supported by NIH grant R01 GM22939. **Author contributions:** S.B. performed simulations, created figures, and led the project. G.B. contributed with theoretical tools for combining simulations with experiments. S.D.K. and D.H.T. collected the experimental data. K.L.-L. and S.B. conceived the study and supervised the project. All authors analyzed data and contributed to writing the paper. **Competing interests:** The authors declare that they have no competing interests. **Data and materials Availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors. Experimental data are available at https://github.com/sbottaro/tetranucleotides_data. The Python script used to reweight trajectories is available at https://github.com/sbottaro/rr.

Submitted 23 December 2017
Accepted 5 April 2018
Published 18 May 2018
10.1126/sciadv.aar8521

Citation: S. Bottaro, G. Bussi, S. D. Kennedy, D. H. Turner, K. Lindorff-Larsen, Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. *Sci. Adv.* **4**, eaar8521 (2018).

Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations

Sandro Bottaro, Giovanni Bussi, Scott D. Kennedy, Douglas H. Turner and Kresten Lindorff-Larsen

Sci Adv **4** (5), eaar8521.
DOI: 10.1126/sciadv.aar8521